arXiv:1508.00310v2 [q-fin.ST] 14 Sep 2015

Statistical Emulators for Pricing and Hedging Longevity Risk Products

J. Risk and M. Ludkovski [1]

Abstract

We propose the use of statistical emulators for the purpose of valuing mortality-linked contracts in stochastic mortality models. Such models typically require (nested) evaluation of expected values of nonlinear functionals of multi-dimensional stochastic processes. Except in the simplest cases, no closed-form expressions are available, necessitating numerical approximation. Rather than building ad hoc analytic approximations, we advocate the use of modern statistical tools from machine learning to generate a flexible, non-parametric surrogate for the true mappings. This method allows performance guarantees regarding approximation accuracy and removes the need for nested simulation. We illustrate our approach with case studies involving (i) a Lee-Carter model with mortality shocks, (ii) index-based static hedging with longevity basis risk; (iii) a Cairns-Blake-Dowd stochastic survival probability model.

Keywords: Statistical emulation, longevity risk, life annuities, valuation of mortality-contingent claims

1. Introduction

Longevity risk has emerged as a key research topic in the past two decades. Since the seminal work of Lee and Carter [28] there has been a particular interest in building stochastic models of mortality. Stochastic mortality allows for generation of a range of future longevity forecasts, and permits the modeler to pinpoint sources of randomness, so as to better quantify respective risk. Longevity modeling calls for a marriage between the statistical problem of calibration, i.e. fitting to past mortality data, and the financial problem of pricing and hedging future longevity risk. At its core, the latter problem reduce to computing expected values of certain functionals of the underlying stochastic processes. For example, the survival probability for t years for an individual currently aged x can be expressed as a functional

$$P(t,x) = \mathbb{E}\left[\exp\left(-\int_0^t \mu(s,x+s)\,ds\right)\right], \tag{1}$$

where $\mu(s,x+s)$ is the force of mortality at date s for an individual aged $x+s$. In the stochastic mortality paradigm $\mu(s,x+s)$ is random for $s>0$, and so one is necessarily confronted with the need to evaluate the corresponding expectations on the right-hand-side of (1).

The past decade has witnessed a strong trend towards complexity in both components of (1). On the one hand, driven by the desire to provide faithful fits (and forecasts) to existing mortality

[1]The authors are with the Department of Statistics & Applied Probability, University of California, Santa Barbara CA 93106-3110;

data, increasingly complex mortality models for $\mu(t,x)$ have been proposed. The latest generation of models feature multi-dimensional, nonlinear stochastic state processes driving $\mu(\cdot,x)$, see e.g. Cairns et al. [11], Li et al. [29], Lin et al. [30], Barrieu et al. [2], Fushimi and Kogure [19]. These models are effective at calibration and emitting desirable forecasts, but lack tractability in terms of closed-form formulas. On the other hand, sophisticated insurance products, such as variable annuities or longevity swap derivatives, make valuation and hedging highly nontrivial, and typically call for numerical approaches, as closed-form formulas are not available. Taken together, pricing of mortality-linked contracts becomes a complex system, feeding multi-dimensional stochastic inputs through a "black box" that eventually outputs net present value of the claim.

These developments have created a tension between the complexity of mortality models that do not admit explicit computations and the need to price, hedge and risk manage complicated contracts based on such models. Due to this challenge, there remains a gap between the academic mortality modeling and the implemented models by the longevity risk practitioners. Because the aforementioned valuation black box is analytically intractable, there is a growing reliance on Monte Carlo simulation tools, which in turn is accompanied by exploding computational needs. For example, many emerging problems require *nested simulations* which can easily take days to complete. Similarly, many portfolios contain millions of heterogeneous products (see, e.g. Gan and Lin [20]) that must be accurately priced and managed. In this article we propose to apply modern statistical methods to address this issue. Our approach is to bridge between the mortality modeling and the desired pricing/hedging needs through an intermediate *statistical emulator*. The emulator provides a computationally efficient, high-fidelity surrogate to the actual mortality model. Moreover, the emulator converts a calibrated opaque mortality model into a user-friendly valuation "app". The resulting toolbox allows a plug-and-play strategy, so that the end user who is in charge of pricing/risk-management can straightforwardly swap one mortality model for another, or one set of mortality parameters for an alternative. This modular approach allows a flexible solution to robustify the model-based longevity risk by facilitating comparisons of different longevity dynamics and different assumptions.

Use of *emulators* is a natural solution to handle complex underlying stochastic simulators and has become commonplace in the simulation and machine learning communities [36, 26]. Below we propose to apply such statistical learning within the novel context of insurance applications. In contrast to traditional (generalized) linear models, emulation calls for fully nonparametric models, which are less familiar to actuaries. To fix ideas, in this article we pursue the problem of pricing/hedging vanilla life annuities, a foundational task in life insurance and pension plan management. Except in the simplest settings, there are no explicit formulas for annuity values and consequently approximation techniques are already commonplace. Looking more broadly, our method would also be applicable in many other actuarial and risk management contexts, see Section 7.

The paper is organized as follows: In Section 2 we introduce the emulation problem and review the mathematical framework of stochastic mortality. Section 3 discusses the construction of emulators, including spline and kriging surrogates, as well as generation of training designs and simulation budgeting. The second half of the paper then presents three extended case studies on several stochastic mortality models that have been put forth in the literature. In Section 4 we examine a Lee-Carter model with mortality shocks that was proposed by Chen and Cox [14]; Section 5 studies approximation of hedge portfolio values in a two-population model based on the recent work by Cairns et al [13]. Lastly, Section 6 considers valuation of deferred annuities under a Cairns-Blake-Dowd (CBD) [9] mortality framework.

2. Emulation Objective

We consider a stochastic system with Markov state process $Z = (Z(t))$. Throughout the paper we will identify Z with the underlying stochastic *mortality factors*. In Section 2.2 we review some of the existing such models and explicit the respective structure of Z. Typically, Z is a multivariate stochastic process based on either a stochastic differential equation or time-series $ARIMA$ frameworks. For example, Z may be of diffusion-type or an auto-regressive process.

In the inference step, the dynamics of Z are calibrated to past mortality data that reflect as closely as possible the population of interest. In the ensuing valuation step, the modeler seeks to evaluate certain quantities related to a functional $F(T, Z(\cdot))$ looking into the future. The time horizon $T \geq 0$ allows consideration of deferred contracts that are common in longevity risk, see below. Our notation furthermore indicates that F potentially depends on the whole path $\{Z(t), t \geq T\}$, such as

$$F(T, Z(\cdot)) = \exp\Big(-\sum_{t=T}^{\infty} h(Z(t))\Big), \tag{2}$$

for some $h(z)$. Given F, the most common aim is to compute its expected value, based on the initial data at $t = 0$,

$$\mathbb{E}\left[F(T, Z(\cdot)) \mid Z(0)\right]. \tag{3}$$

Other summary statistics of interest in actuarial applications include the

- Quantile $q(\alpha; F(T, Z(\cdot)))$ (eg. the Value-at-Risk at level α of F);
- Expected Shortfall of F, $\mathbb{E}[F(T, Z(\cdot)) \mid F(T, Z(\cdot)) \leq q(\alpha; F(T, Z(\cdot))), Z(0)]$;
- Correlation between two functionals, $Corr(F_1(T, Z(\cdot)), F_2(T, Z(\cdot))|Z(0))$.

To fix ideas we henceforth focus on (3) which is a fundamental quantity in pricing/hedging problems. When $T > 0$, the evaluation of (3) can be broken into two steps, namely first we evaluate

$$f(z) \doteq \mathbb{E}[F(T, Z(\cdot))|Z(T) = z], \tag{4}$$

and then use the Markov property of Z to carry out an outer average,

$$\mathbb{E}[F(T, Z(\cdot))|Z(0)] = \int_{\mathbb{R}^d} f(z) p_T(z|Z(0)) dz,$$

where $p_T(z'|z) = P(Z(T) = z'|Z(0) = z)$ is the transition density of Z over $[0, T]$.

Crucially, because the form of $F(T, Z(\cdot))$ is nontrivial, we shall assume that $f(z)$ is not available explicitly, and there is no simple way to describe its functional form. However, since $f(z)$ is a conditional expectation, it can be sampled using a simulator, i.e. the modeler has access to an engine that can generate independent, identically distributed samples $F(T, Z^{(n)}(\cdot))$, $n = 1, \ldots,$ given $Z(0)$. However this simulator is assumed to be expensive, implying that computational efficiency is desired in using it.

Given an initial state $Z(0)$, a naive Monte Carlo approach to evaluate (3) is based on nested simulation. First, the outer integral over $p_T(z|Z(0))$ is replaced by an empirical average of (4) across $m = 1, \ldots, N_{out}$ draws $z^{(m)} \sim Z(T)|Z(0)$,

$$\mathbb{E}[F(T, Z(\cdot))|Z(0)] \simeq \frac{1}{N_{out}} \sum_{m=1}^{N_{out}} f(z^{(m)}). \tag{5}$$

Second, for each $z^{(m)}$ the corresponding inner expected value $f(z^{(m)})$ is further approximated via

$$f(z^{(m)}) \simeq \frac{1}{N_{in}} \sum_{n=1}^{N_{in}} F(T, z^{(m),n}(\cdot)), \qquad m = 1, \ldots, N_{out}, \tag{6}$$

where $z^{(m),n}(t), t \geq T$ are N_{in} independent trajectories of Z with a fixed starting point $z^{(m),n}(T) = z^{(m)}$. This nested approach offers an unbiased but expensive estimate. Indeed, the total simulation budget is $\mathcal{O}(N_{out} \cdot N_{in})$ (where the usual big-Oh notation $h(x) = \mathcal{O}(x)$ means that $h(\cdot)$ is asymptotically linear in x as $x \to \infty$) which can be computationally intensive – for example 1,000 simulations at both steps requires 10^6 total simulations.

For this reason, it is desirable to construct cheaper versions of approximating (3). The main idea is to replace the inner step of repeatedly evaluating $f(z)$ (possibly for some very similar values of z) with a simpler alternative. One strategy is to construct deterministic approximations to (4) by replacing the random variable $Z(s)|Z(T)$, $s > T$ with a fixed constant, e.g. its mean, which can then be plugged into F to estimate the latter's expected value. This effectively removes the stochastic aspect and allows to obtain explicit approximations to $f(\cdot)$. (The simplest approximation is to simply freeze $Z(s) = Z(T) \forall s > T$.) However, the resulting error is hard to judge, and moreover, analytic, off-line derivations are needed to obtain a good approximation. Consequently, we advocate the more statistical method of utilizing a *surrogate* model for $f(\cdot)$. This approach can be generically used in any Markovian setting, requires no analytic derivations, and makes minimal a priori assumptions about the structure of $f(\cdot)$.

The main idea of emulation is a regression framework that generates a fitted $\hat{f}(\cdot)$ by solving regression equations over a training dataset $\{z^{(n)}, F(T, z^{(n)}(\cdot))\}_{n=1}^{N_{tr}}$ of size N_{tr}. Emulation reduces approximating $f(\cdot)$ to the twin statistical problems of (i) experimental design (generating the training dataset) and (ii) regression (specifying the optimization problem that the approximation $\hat{f}$ solves). Details of these steps are presented in Section 3 below.

Because we are fitting a full response model, rather than a pointwise estimate, the emulator budget $N_{tr} \gg N_{in}$ will be an order of magnitude bigger than in (6). It will also require regression overhead. However, once $\hat{f}$ is fitted, prediction of $\hat{f}(z)$ for a particular value z takes $\mathcal{O}(1)$ effort, so that we can use (5) to estimate the original problem in (3) at a cost linear in N_{out}. To sum up, the total budget of the emulator is just $\mathcal{O}(N_{tr} + N_{out})$, much smaller than $\mathcal{O}(N_{out} \times N_{in})$ of nested Monte Carlo. These savings become even more significant as the dimension of state Z grows. Indeed, with multi-dimensional models, both N_{out} and N_{in} need to be larger to better cover the respective integrals over $\mathbb{R}^d$, and hence the efficiency of nested simulations will deteriorate quickly. Intuitively, the latter computational budget is at least quadratic in d. In contrast, the intuitive complexity of an emulator is linear in d. As stochastic mortality models become more complex, models with $d = 3, 4, 5+$ factors are frequently proposed, and efficiency issues become central to the ability of evaluating (3) tractably.

2.1. Valuation of Life Annuities

In longevity modeling, Z represents the stochastic factors driving the central force of mortality $m(t,x)$. Formally, $Z = (Z(t)) = (Z_1(t),\ldots,Z_d(t))$ is a $d-$dimensional $(\mathcal{F}(t))$ measurable Markov process on a complete filtered probability space $(\Omega,\mathcal{F},\mathbb{P},(\mathcal{F}(t)))$. The filtration $(\mathcal{F}(t))$ is the information up to time t of the evolution of the mortality processes.

A typical state-of-the-art model decomposes $m(t,x)$ into a longevity trend, an Age effect, and a Cohort effect (known collectively as APC models). Each of the above may be modeled in turn by one or more stochastic factors. The most common models are the Lee-Carter [28] and CBD [9] models and their generalizations. Generally their individual components follow an ARIMA model; details can be found in the survey Cairns et al. [11].

To deal with cashflows at different dates, we assume the existence of a risk-free asset and denote by $B(T,T+s)$ the price of an s-bond at date T with maturity at $T+s$. For the rest of the article we will assume constant force of interest r, leading to $B(T,T+s) = e^{-rs}$. One can straightforwardly handle stochastic interest rates (which then form part of $Z(\cdot)$); see Jalen and Mamon [25] for a discussion of correlation structure between mortality and interest rates and Fushimi and Kogure [19] for an example that applies Bayesian methods to longevity derivative pricing under a Cox-Ingersoll-Ross interest rate model.

Consider an individual aged x at time 0 whose remaining lifetime random variable is denoted as τ_x. The state process Z captures $m(t,x+t)$, the mortality rate process for τ_x at time t, when the individual would have aged to $x+t$. For small dt, the instantaneous probability of death is approximately $m(t,x+t)dt$, so that the random survival function of τ_x is

$$S(t,x) \doteq \exp\left(-\int_0^t m(s,x+s)\right). \tag{7}$$

More generally for $u \le t < T$, the probability of an individual aged x to survive between dates t and T, given the information at time u is given by

$$\begin{aligned}\mathbb{P}(\tau_x > T \mid \tau_x > t,\mathcal{F}_u) &= \mathbb{E}\left[\frac{S(T,x)}{S(t,x)}\middle|\,\mathcal{F}_u\right] \\ &= \mathbb{E}\left[\exp\left(-\int_t^T m(s,x+s)\right)\middle|\,Z(u)\right] \doteq P(Z(u);t,T,x),\end{aligned} \tag{8}$$

where the last equality follows from the Markov property. The deterministic analogue of $P(Z(0);t,T,x)$ in actuarial literature is ${}_{T-t}p_{x+t}$.

As a canonical actuarial contract, we henceforth focus on deferred life annuities. These contracts are fundamental to valuation of defined benefit pension plans, which normally begin paying annuitants at retirement age (typically age 65) and continue until their death, possibly with survivor benefits. (For valuation purposes the payment is assumed to end at some pre-specified upper age $\bar{x}$, e.g. 100 or 110). A major problem of interest is valuing such life annuities for current plan participants who are still working, i.e. under age 65. Because this requires making longevity projections many decades into the future, longevity risk becomes a crucial part of risk management. The net present value of a life annuity at date T is

$$a(Z(T);T,x) \doteq \sum_{s=1}^{\infty} B(T,T+s)\mathbb{P}(\tau_x \ge T+s \mid \mathcal{F}_T) = \sum_{s=1}^{\bar{x}-x} e^{-rs} P(Z(T);T,T+s,x), \tag{9}$$

where we emphasize that the random mortality shocks come from Z. Finally, the net present value at $t = 0$ is $NPV \doteq \mathbb{E}[e^{-rT} \cdot a(Z(T), T, x)]$, which can be seen as an instance of (3) that includes discounting and integrating over the density of $Z(T)$. Except for the simplest models, the survival probability $P(z; \cdot)$ is not analytically known and hence neither is (9) or NPV. Without a representation for $z \mapsto a(z, T, x)$ one is then forced to resort to approximations for all the basic tasks of pricing, hedging, asset liability management, or solvency capital computation. The discussed nested simulation takes the form of first approximating $a(z^{(1)}, T, x)$ for some representative scenarios $(z^{(1)}, \ldots, z^{(n)})$, and then further manipulating the resulting "empirical" distribution of $(a(z^{(1)}, T, x), \ldots, a(z^{(n)}, T, x))$. Emulation provides a principled statistical framework for optimizing, assessing and improving such two-level simulations.

Remark. As mentioned, estimation of $a(\cdot, T, x)$ is usually a building block embedded in a larger setting which requires repeated evaluation of the former quantity. For instance, Bauer et al. [4] addresses nested Monte Carlo simulations in calculating the present value of life-annuity-like instruments in the calculation of solvency capital requirements. Let us also mention the works Bacinello et al. [1], Boyer and Stentoft [6] who considered valuation of mortality contracts with early exercise features, such as surrender guarantees. The respective least squares Monte Carlo algorithms can be seen as classical parametric linear-model emulators in our terminology, eschewing the need for nested Monte Carlo forests.

2.2. Stochastic Mortality

We concentrate on discrete-time mortality models which are easier to calibrate to the discrete mortality data, typically aggregated into annual intervals. The common assumption is that the central force of mortality remains constant through a given calendar year, so that for all $0 \le s, u \le 1$, we have $m(t+s, x+u) = m(t, x)$. Therefore

$$P(Z(u); t, T, x) = \mathbb{E}\left[\exp\left(- \sum_{s=t+1}^{T} m(s, x+s) \right) \middle| Z(u) \right], \quad u \le t < T. \tag{10}$$

Thus, $P(Z(T); T, T+s, x+T)$ becomes a functional of the trajectory of Z.

Three major approaches to stochastic mortality have been put forward in the literature. The first approach, pioneered by Lee and Carter [28], directly treats $m(t, x)$ as a product of individual stochastic processes, e.g. $ARIMA$ time-series. This setup allows incorporating demographic insights, as well as disentangling age, period and cohort effects in future forecasts. To wit, the popular age-period-cohort (APC) mortality models assume that (see Appendix A for more details)

$$\log m(t, x) = \beta_x^{(1)} + \frac{1}{n_a}\kappa^{(2)}(t) + \frac{1}{n_a}\gamma^{(3)}(t-x), \tag{11}$$

where $\kappa^{(2)}$ and $\gamma^{(3)}$ are stochastic processes and n_a is the number of ages that x can take in fitting. In this case, the state process $Z(t)$ depends on current and potentially past values of $\kappa^{(2)}$ and $\gamma^{(3)}$. Attempts to understand the statistical validity of such models have been done by, for example, Lee and Miller [27], [8], Booth et al. [5], Czado et al. [16], Delwarde et al. [18], and Li et al. [29]. In addition, there have been several extensions of the Lee Carter model by Renshaw and Haberman [34], Hyndman and Ullah [24], Plat [32], Debonneuil [17], and Cairns et al. [12].

None of these models admit closed form expressions for survival probabilities $P(z; \cdot)$. Consequently, several authors have proposed approximation methods. Coughlan et al. [15] used a

bootstrapping approach, while Cairns et al. [13] derived an analytic approximation, commenting that industry practice is to utilize deterministic projections. The more flexible tool of Monte Carlo simulation has been applied in Bauer et al. [4] among others.

The second approach, due to Cairns et al. [9] (CBD), generates a stochastic model for the survival probability (10), allowing for straightforward pricing of longevity-linked products; however, it is more difficult to calibrate and to obtain reasonable forecasts for future mortality experience in a population as a whole. The third approach works with forward mortality rates [3], borrowing ideas from fixed income markets. Forward models give a holistic view of how the mortality curves can evolve over time, and presents a dynamically consistent structure for mortality forecasting. Once again however, they do not provide easy expressions for (10) and hence require further manipulation for pricing purposes.

2.3. Bias/Variance Trade-Off

With a view towards approximating (9), it is imperative to first quantify the resulting quality of an approximation. The standard statistical approach is to use the framework of mean squared error. Fix z and let $a(z) \equiv a(z,T,x)$ be the true value of a life annuity conditional on state $Z(T) = z$. If $a(z)$ is being estimated by $\hat{a}(z)$, then

$$\mathrm{IMSE}(\hat{a}) \doteq \mathbb{E}\left[(\hat{a}-a)^2\right], \qquad \mathrm{Bias}(\hat{a}) = \mathbb{E}\left[\hat{a}-a\right], \tag{12}$$

where the averaging is over the sampling distribution (i.e. different realizations of data used in constructing it) of $\hat{a}(z)$.

Starting with (12) leads to the fundamental bias/variance trade-off. At one end of the spectrum, a Monte Carlo estimate as in (6) has zero bias but carries a high variance. At the opposite end, an analytic approximation has zero variance, but will have a non-zero bias that cannot be alleviated (whereas the Monte Carlo IMSE will go to zero as the size of the dataset grows $N_{tr} \to \infty$) even asymptotically. Because low variance is often preferred practically, analytic methods have remained popular. Cairns et al. [13] echoes that it is usual practice in industry to use a deterministic projection of mortality rates rather than use a simulation approach.

The basic idea for the deterministic approximations is that if $\hat{m}(t,x)$ is an unbiased estimate for $m(t,x)$, then

$$\begin{aligned} P(Z(u);t,T,x) &= \mathbb{E}\left[\exp\left(-\sum_{s=t+1}^{T} m(s,x+s)\right)\middle|\, Z(u)\right] \\ &\approx \exp\left(-\sum_{s=t+1}^{T}\mathbb{E}\left[m(s,x+s)\mid Z(u)\right]\right) = \exp\left(-\sum_{s=t+1}^{T}\hat{m}(s,x+s;Z(u))\right). \end{aligned} \tag{13}$$

Using the estimate for $P(Z(u);\cdot)$ in (13) one can then approximate $a(Z(T),T,x)$ term-wise. Jensen's inequality implies that $\exp\left(-\sum_{s=t+1}^{T}\hat{m}(s,x+s)\right) > P(Z(u);t,T,x)$. Consequently, any such approximation is guaranteed to be biased high for the survival probabilities (and subsequently the annuity values).

Analytic approximations can be very powerful and of course very fast but they carry two major disadvantages. One is the need to *derive* a suitable estimator $\hat{m}$. This may be possible in a simple model (e.g. low-dimensional Z with linear dynamics, like in the original Lee-Carter model), but otherwise may require a lot of off-line labor, leading to unnecessary focus on simplifications at the

expense of calibration and risk management consistency. Second, the degree of accuracy of the approximation is unknown. Indeed, there is generally not much that is available about empirical accuracy of the right-hand-side in (13) for a given model, leaving the user in the dark about how much error is being made. This issue is very dangerous, since potentially major mis-valuations may creep up unbeknownst to the risk manager.

To remedy the above shortcomings, while still maintaining significant variance reduction compared to plain MC, we advocate the use of statistical emulators. The latter offer posterior quantification of accuracy (via standard error or Bayesian posterior variance), and do not require any simplifications of the mortality model. An additional advantage is that one can directly approximate $z \mapsto a(z, T, x)$ without having to do intermediate approximations of the survival probabilities (which inevitably lead to further error compounding). As we demonstrate in the case studies, statistical models for $a(z)$ can indeed efficiently address the bias/variance trade-off, by maintaining negligible bias and small variance, leading to improved IMSE metrics compared to other approaches.

3. Statistical Emulation

The idea of emulation is to replace the computationally expensive process of running a Monte Carlo sub-routine to evaluate $f(z)$ for each new site z with a cheap-to-evaluate surrogate model that statistically predicts $f(z)$ for any $z \in \mathbb{R}^d$ based on results from a training dataset. At the heart of emulation is statistical learning. Namely, the above predictions are based on first obtaining pathwise estimates $y^{(n)} = F(T, z^{(n)})$, $n = 1, \ldots, N_{tr}$ for a set of training locations, called a design $\mathcal{D} \doteq (z^{(1)}, \ldots, z^{(N_{tr})})$. Next, one regresses $\{y^{(n)}\}$ against $\{z^{(n)}\}$ to "learn" the response surface $\hat{f}(\cdot)$. The regression aspect allows to borrow information across different scenarios starting at various sites. This reduces computational budget compared to the nested simulation step of independently making N_{tr} pointwise estimates $f(z^{(n)})$ by running N_{in} scenarios from *each* site $z^{(n)}$. The conceptual need for regression is two-fold. First, the emulator is used for interpolation, i.e. using existing design to make predictions at new sites z. In contrast, plain Monte Carlo only predicts at $z^{(n)}$'s. Second, like in the classical approach, the emulator *smoothes* the Monte Carlo noise from sampling trajectories of $\{Z(s), s > T\}$.

Formally, the statistical problem of emulation deals with a sampler (or oracle)

$$Y(z) = f(z) + \epsilon(z), \tag{14}$$

where we identify $f(z) \equiv a(z, T, x)$ with the unknown *response surface* and ϵ is the sampling noise, assumed to be independent and identically distributed across different calls to the oracle. We make the assumption $\epsilon(z) \sim N(0, \tau^2(z))$, where $\tau^2(z)$ is the sampling variance that depends on the location z. Emulation now involves the (i) experimental design step of proposing a design $\mathcal{D}$ that forms the training dataset, and (ii) a learning procedure that uses the queried results $(z^{(n)}, y^{(n)})_{n=1}^{N_{tr}}$, with the $y^{(n)}$ being realizations of (14) given $z^{(n)}$, to construct a fitted response surface $\hat{f}(\cdot)$. The fitting is done by specifying the approximation function class $\hat{f} \in \mathcal{H}$, and a loss function $L(\hat{f}, f)$ which is to be minimized. The loss function measures the relative accuracy of $\hat{f}$ vis-a-vis the ground truth; in this paper we focus on the mean-squared approximation error

$$L(\hat{f}, f) \doteq \int_{\mathbb{R}^d} |\hat{f}(z) - f(z)|^2 dz. \tag{15}$$

Because the true f is unknown, the definition of $L(\hat{f}, f)$ cannot be operationalized and instead a proxy based on the uncertainty (such as Bayesian posterior uncertainty or standard errors)

surrounding $\hat{f}$ is applied. Also, since the structure of f is unknown, it is desirable that the approximation class $\mathcal{H}$ is dense, i.e. has a sufficiently rich architecture to approximate any f to an arbitrary degree of accuracy. To this end, we concentrate on kernel regression methods, namely linear smoothers. In the next subsections we introduce two such regression families, smoothing splines and kriging (Gaussian process) models.

Remark. In this paper we focus on the original task of producing an accurate approximation to f everywhere. In some contexts, accuracy is judged not globally, but locally, so that a differentiated accuracy measure is used. For example, in VaR applications, the model for f must be accurate in the left-tail, but can be rather rough in the right-tail. In this case, (15) can be replaced by a weighted loss metric.

3.1. Emulators based on Spline Models

We generate emulators $\hat{f}(\cdot)$ using a regularized regression criterion. To wit, given a smoothing parameter $\lambda \geq 0$ we look for the minimizer $\hat{f} \in \mathcal{H}$ of the following penalized residual sum of squares problem

$$RSS(f, \lambda) = \sum_{n=1}^{N_{tr}} \{y^{(n)} - f(z^{(n)})\}^2 + \lambda J(f), \tag{16}$$

where $J(f)$ is a penalty or regularization function. We concentrate on the case where the approximation class has a reproducing kernel Hilbert space (RKHS) structure which also generates $J(f)$. Namely, there exists an underlying positive definite kernel $C(z, z')$ such that $\mathcal{H}_C = \text{span}(C(\cdot, z) : z \in \mathbb{R}^d)$ is the Hilbert space generated by C and $J(f) = \|f\|^2_{\mathcal{H}_C}$. The representer theorem implies that the minimizer of (16) has an expansion in terms of the eigen-functions

$$\hat{f}(z) = \sum_{j=1}^{N_{tr}} \alpha_j C(z, z^{(j)}), \tag{17}$$

relating the prediction at z to the kernel function sampled at the design sites $z^{(j)}$.

Our first family are smoothing (or thin-plate) splines that take

$$J(f) = \int_{\mathbb{R}^d} \left[\sum_{i,j=1}^{d} \frac{\partial}{\partial z_i} \frac{\partial}{\partial z_j} f(z) \right] dz, \tag{18}$$

and $\mathcal{H}$ as the set of all twice continuously-differentiable functions. It is known [21, Chapter 5] that in this case the underlying kernel is given by $C(z, z') = \|z - z'\|^2 \log \|z - z'\|$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$. The resulting optimization of (16) along with (18) gives a smooth response surface which is called a thin-plate spline (TPS), and has the explicit form

$$f(z) = \beta_0 + \boldsymbol{\beta}^T \vec{z} + \sum_{j=1}^{N_{tr}} \alpha_j \|z - z^{(j)}\|^2 \log \|z - z^{(j)}\|, \tag{19}$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$.

In 1-d, the penalized optimization reduces to

$$\inf_{f\in\mathcal{C}^2}\sum_{i=1}^{N_{tr}}\{y^{(n)}-f(z^{(n)})\}^2+\lambda\int_{\mathbb{R}}\{f''(u)\}^2du. \tag{20}$$

The summation in (20) is a measure of closeness of data, while the integral penalizes the fluctuations of f. Note that $\lambda=\infty$ reduces to the traditional least squares linear fit $\hat{f}(z)=\beta_0+\beta_1 z$ since it introduces the constraint $f''(z)=0$. It is well known that the resulting solution is an expansion in terms of natural cubic splines, i.e. $\hat{f}$ is a piecewise cubic polynomial that has continuous first and second derivatives at the design sites $z^{(n)}$, and is linear outside of the design boundary.

Several methods are available to choose the smoothing parameter λ, including cross-validation or MLE Hastie et al. [21, Chapter 5]. A common parametrization is through the effective degrees of freedom statistic df_λ. We use the `R` package "fields" [31] to fit multi-dimensional thin plate splines, and the base `smooth.spline` function for the one-dimensional case.

3.2. Kriging Surrogates

A kriging surrogate assumes that f in (14) has the form

$$f(z)=\mu(z)+X(z), \tag{21}$$

where $\mu:\mathbb{R}^d\to\mathbb{R}$ is a trend function, and X is a mean-zero square-integrable process. Specifically, X is assumed to be a realization of a Gaussian process with covariance kernel C. The role of C is identical to the regularized regression above, i.e. C generates the approximating family $\mathcal{H}_C$ that X is assumed to belong to.

However, kriging also brings a Bayesian perspective, treating X as a random function to be learned, and estimation as computing the posterior distribution of X given the collected data $\mathbf{y}\doteq(y^{(1)},\dots,y^{(N_{tr})})$. The RKHS framework implies that the posterior mean (more precisely its maximum a posteriori estimate) of $X(z)$ coincides with the regularized regression prediction from the previous section. In the Bayesian framework, C is interpreted as the covariance kernel, $C(z,z')=Cov(f(z),f(z'))$ as $f(\cdot)$ ranges over $\mathcal{H}_C$. Assuming that the noise $\epsilon(z)$ is also Gaussian implies that $X(z)|\mathbf{y}\sim N(m(z),s^2(z))$ has a Gaussian posterior, which reduces to computing the kriging mean $m(z)$ and kriging variance $s^2(z)$.

In turn, the kriging variance $s^2(z)$ offers a principled empirical estimate of model accuracy, quantifying the approximation quality. In particular, one can use $s^2(z)$ as the proxy for the MSE of $\hat{f}$ at z. Integrating $s^2(z^{(n)})$ over the outer design locations then yields an assessment regarding the error of (3).

3.2.1. Simple Kriging

Simple kriging (SK) assumes that the trend $\mu(z)$ is known. By considering the process $f(z)-\mu(z)$, we may assume without loss of generality that $f(z)$ is centered at zero and $\mu\equiv 0$. The resulting posterior mean and variance are then [35]

$$\begin{cases} m_{SK}(z)\doteq\mathbf{c}(z)^T\mathbf{C}^{-1}\mathbf{y};\\ s^2_{SK}(z)\doteq C(z,z)-\mathbf{c}(z)^T\mathbf{C}^{-1}\mathbf{c}(z),\end{cases} \tag{22}$$

where $\mathbf{c}(z)=\big(C(z,z^{(n)})\big)_{1\le n\le N_{tr}}$ and

$$\mathbf{C}\doteq\Big[C(z^{(i)},z^{(j)})\Big]_{1\le i,j\le N_{tr}}+\mathbf{\Delta}, \tag{23}$$

with $\mathbf{\Delta}$ the diagonal matrix with entries $\tau^2(z^{(1)}), \ldots, \tau^2(z^{(N_{tr})})$.

3.2.2. Universal Kriging

Universal kriging (UK) generalizes (21) to the case of a parametric trend function of the form $\mu(z) = \beta_0 + \sum_{j=1}^p \beta_j h_j(z)$ where β_j are constants to be estimated, and $h_j(\cdot)$ are given basis functions. The coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is estimated simultaneously with the Gaussian process component $X(z)$. A common choice is first-order UK that uses $h_j(z) = z_j$ for $j = 1, \ldots, d$. Another common choice is zero-order UK, also known as Ordinary Kriging (OK) that takes $\mu(z) = \beta_0$ a constant to be estimated.

If we let $\mathbf{h}(z) \doteq (h_1(z), \ldots, h_p(z))$ and $\mathbf{H} \doteq \left(\mathbf{h}(z^{(1)}), \ldots, \mathbf{h}(z^{(N)})\right)$, then the universal kriging mean and variance at location z are [35]

$$\begin{cases} m_{UK}(z) = \mathbf{h}(z)^T \hat{\boldsymbol{\beta}} + \mathbf{c}(z)^T \mathbf{C}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}); \\ s^2_{UK}(z) = s^2_{SK}(z) + \left(\mathbf{h}(z)^T - \mathbf{c}(z)^T\mathbf{C}^{-1}\mathbf{H}\right)^T \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1} \left(\mathbf{h}(z)^T - \mathbf{c}(z)^T\mathbf{C}^{-1}\mathbf{H}\right), \end{cases} \tag{24}$$

where the best linear estimator of the trend coefficients $\boldsymbol{\beta}$ is given by the usual linear regression formula $\hat{\boldsymbol{\beta}} \doteq \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1} \mathbf{H}^T\mathbf{C}^{-1}\mathbf{y}$.

The combination of trend and Gaussian process (GP) model offers an attractive framework for fitting a response surface. The trend component allows to incorporate domain knowledge about the response, while the GP component offers a flexible nonparametric correction. One strategy is to specify a known trend (coming from some analytic approximation) and fit a GP to the residuals, yielding a Simple Kriging setup. Another strategy is to take a low-dimensional parametric approximation, such as a linear function of Z-components, and again fit a GP to the residuals, leading to a Universal Kriging setup.

3.2.3. Covariance kernels and parameter estimation

The covariance function $C(\cdot,\cdot)$ is a crucial part of a Kriging model. In practice, one usually considers spatially stationary or isotropic kernels,

$$C(z, z') \equiv c(z - z') = \sigma^2 \prod_{j=1}^d g((z - z')_j; \theta_j),$$

reducing to the one-dimensional base kernel g. Below we use the power exponential kernels $g(h;\theta) = \exp\left(-\left(\frac{|h|}{\theta}\right)^p\right)$. The hyper-parameters θ_j are called characteristic length-scales and can be informally viewed as roughly the distance you move in the input space before the response function can change significantly, Rasmussen and Williams [33, Ch 2]. The user-specified power $p \in [1, 2]$ is usually taken to be either $p = 1$ (the exponential kernel) or $p = 2$ (the Gaussian kernel). Fitting a kriging model requires picking a kernel family and the hyper-parameters σ_j, θ_j. Two common estimation methods are maximum likelihood, using the likelihood function based on the distributions described above, and penalized MLE (PMLE). Either case leads to a nonlinear optimization problem to fit θ_j and process variance σ^2. One can also consider Bayesian Kriging, where trend and/or covariance parameters have a prior distribution, see Helbert et al. [22]. We utilize the `R` package "DiceKriging" [35] that allows fitting of SK and UK models with five options for a covariance kernel family, and several options on how the hyper-parameters are to be estimated.

3.2.4. Batching

To construct an accurate emulator for $f(\cdot)$, it is important to have a good estimate of the sampling noise $\tau^2(z)$. Typically this information is not available to the modeler a priori. One of the advantages of plain nested Monte Carlo is that generating N_{in} scenarios from a fixed $z^{(n)}$ gives natural empirical estimates *both* for $f(z^{(n)})$ and $\tau^2(z^{(n)})$. To mimic this feature, we therefore consider batched or replicated designs $\mathcal{D}$. To wit, given a total budget of $N_{tr} = N_{tr,1} \cdot N_{tr,2}$ training samples, we allocate them into $N_{tr,1}$ distinct design sites $z^{(1)}, \dots, z^{(N_{tr,1})}$, and then generate $N_{tr,2}$ trajectories from each $z^{(n)}$. Next, the above batches are aggregated into

$$y^{(n)} \doteq \frac{1}{N_{tr,2}} \sum_{j=1}^{N_{tr,2}} F(T, z^{(n),j}(\cdot)); \tag{25}$$

$$\hat{\tau}^2(z^{(n)}) \doteq \frac{1}{N_{tr,2}-1} \sum_{j=1}^{N_{tr,2}} \left\{ y^{(n)} - F(T, z^{(n),j}(\cdot)) \right\}^2, \tag{26}$$

and the resulting dataset $\{z^{(n)}, y^{(n)}, \hat{\tau}^2(z^{(n)})\}$, $n = 1, \dots, N_{tr,1}$ is used to fit a kriging model for $\hat{f}$, with $\hat{\tau}^2(z^{(n)})/N_{tr,2}$ proxying the simulation variance at $z^{(n)}$.

The efficient allocation between $N_{tr,1}$ and $N_{tr,2}$ was analyzed in Broadie et al. [7] for a related risk management problem and it was shown that the optimal choices satisfy

$$N_{tr,1} \propto N_{tr}^{2/3}, \qquad N_{tr,2} \propto N_{tr}^{1/3}. \tag{27}$$

This is also the allocation we pursue in this paper, so that there are relatively many more design sites than replications in each batch.

3.3. Experimental Design

Several approaches are possible for constructing the training design $\mathcal{D}$. First, one may generate an empirical design by independently sampling $z^{(n)} \sim Z(T)|Z(0)$. This allows to emulate the conditional density $p_T(z|Z(0))$ which is advantageous for computing an expectation like in (3). Second, one may generate a random $\mathcal{D}$ using some other proposal density $z^{(n)} \sim Q$. For example, a uniform proposal density (i.e. $z^{(n)}$ i.i.d. uniform in some domain $D \subseteq \mathbb{R}^d$) yields a basic attempt in having a space filling experimental design of arbitrary size. A more structured (but still random) design can be obtained via Latin Hypercube Sampling (LHS) techniques [38]. Roughly speaking, LHS builds a regular d-dimensional lattice and then attempts to equidistribute $N_{tr,1}$ sites among the resulting hypercubes. Within each selected hypercube the design site is placed uniformly.

Third, one can use a deterministic design, such as a latticed grid, or a quasi-Monte Carlo (QMC) sequence. Deterministic designs ensure a space-filling property and easy reproducibility. For example, the Sobol sequence [37] redistributes a uniform binary grid to produce a grid that is maximally equidistributed. Compared to LHS, use of QMC is faster (as it can be directly hard-coded) and can be manually tweaked as needed. Both methods reduce Monte Carlo variance of $\hat{f}$ relative to empirical $\mathcal{D}$. Theoretically, the typical domain of $Z(T)$ is unbounded, e.g. $\mathbb{R}+^d$. This is not an issue for empirical design construction; for LHS and QMC methods, one must specify an appropriate bounding domain $D \in \mathbb{R}^d$ before generating $\mathcal{D}$.

Remark. Depending on the context, the design $\mathcal{D}$ might need to be spatially non-uniform. For example, if using a deterministic design for computing (3), it may be preferable to capture the

correlation structure among the components of $Z(T)$, or to up-weigh the regions most likely for $Z(T)$. If one is estimating a quantile or tail expectation, $\mathcal{D}$ should preferentially cover the extreme values of the distribution of $Z(T)$; in that situation, an empirical design would be inappropriate.

3.3.1. Generating Longevity Scenarios

Construction of an emulator entails the basic building block of generating a longevity scenario $\{Z(t), t = 0, \dots\}$. In the simplest setting, this just requires to generate and manipulate a sequence of i.i.d Uniform draws that describe the random increments of the (components) of Z. However, typically the model used also includes parameters that must be estimated or calibrated. This aspect becomes nontrivial when future longevity projections are made, whereby model re-fitting may be carried out. Re-fitting introduces path-dependency, making parameters dynamic quantities that might need to be included in Z. For example, Cairns et al. [13] advocate the PPC (partial parameter certain) scenario generation that breaks the overall simulation into two pieces of $[0, T]$ and $[T, \infty)$. With PPC, one initially calibrates the model at $t = 0$ using past mortality data and then simulates up to time T. The simulated scenario is then appended to the historical data, so that the simulation becomes the new "history" from time 0 to time T. The model parameters are then re-fitted at T and the resulting, modified longevity dynamics of Z are used to simulate beyond T. The idea of PPC is to capture some memory of mortality evolution, in essence removing some of the presumed Markovian structure. Under PPC the refitted parameters are blended into $Z(T)$ since they affect the resulting $F(T, Z(\cdot))$.

Also, in the interest of dimension reduction, one could drop some components of the full state space when constructing the emulator. To do so, one may analyze what dynamic variables materially impact annuity values, for example via some simple regression models to test for statistical significance.

3.4. Fitting and Evaluation of Emulators

To fit an emulator for a given simulation budget N_{tr}, we first decompose $N_{tr} = N_{tr,1} \times N_{tr,2}$ and then construct an experimental design $\mathcal{D}$ of size $N_{tr,1}$ using one of the methods in Section 3.3. Each site in $\mathcal{D}$ then spawns $N_{tr,2}$ trajectories that are batched together as in (25). Fitting is done in `R` using the mentioned publicly available packages. For kriging, we use the default setting of the `km` function in DiceKriging package.

Given $\hat{a}(z, T, x)$ we evaluate its performance across a test set $\mathcal{D}^{test} = (z^{(1)}, \dots, z^{(N_{out})})$ of N_{out} locations. Note that $\mathcal{D}^{test}$ is distinct from the training set $\mathcal{D}$. In line with (3) we use an empirical testing set $\mathcal{D}^{test}$: $z^{(n)} \sim Z(T)|Z(0)$. Since the true values $a(z, T, x)$ are not available, we benchmark against an (expensive) gold standard estimate $\hat{a}^{MC}(z, T, x)$ that is described below. In particular, we record the integrated MSE and Bias statistics from (12), namely

$$\widehat{\text{IMSE}}\,(\hat{a}) = \frac{1}{N_{out}} \sum_{n=1}^{N_{out}} \Big(\hat{a}(z^{(n)}, T, x) - \hat{a}^{MC}(z^{(n)}, T, x) \Big)^2; \tag{28}$$

$$\widehat{\text{Bias}}\,(\hat{a}) = \frac{1}{N_{out}} \sum_{n=1}^{N_{out}} \Big[\hat{a}(z^{(n)}, T, x) - \hat{a}^{MC}(z^{(n)}, T, x) \Big]. \tag{29}$$

For benchmarking, we use a high-fidelity nested Monte Carlo approach (5)-(6). While expensive, it is a simple, asymptotically consistent, unbiased estimator. Specifically, for valuing annuities, $\hat{a}^{MC}(z, T, x)$ is obtained by averaging $N_{in} = 1000$ scenarios of $\{Z(s), s > T\}$ at each $z^{(n)} \in \mathcal{D}^{test}$.

Unless indicated otherwise, we use $N_{out} = 1000$, so that the overall budget of $\hat{a}^{MC}$ is $\mathcal{O}(N_{out} \times N_{out})$. We then compare against emulators that use $N_{tr} \in [100, 8000]$, which yields an efficiency gain on the order of 10-50x speed-up. We also compare against deterministic estimators that require no training at all (but do need an analytic derivation), and take just $\mathcal{O}(N_{out})$ budget to make predictions for the outer N_{out} simulations to evaluate (28).

4. Case Study: Predicting Annuity Values under a Lee-Carter with Shocks Framework

Chen and Cox [14] introduced a mortality model based on the traditional Lee Carter set-up:

$$\log m(t, x) = \beta^{(1)}(x) + \beta^{(2)}(x)\kappa^{(2)}(t). \tag{30}$$

This is the same as the APC model (M2) in Appendix A without the cohort term. In the Chen-Cox model, $\beta^{(1)}(x)$ and $\beta^{(2)}(x)$ are deterministic vectors capturing age effects, and $\kappa^{(2)}(t)$ is a stochastic process capturing the period effect with dynamics

$$\kappa^{(2)}(t+1) = \kappa^{(2)}(t) + \mu^{(1)} + \xi^{(1)}(t+1) + [\xi^{(2)}(t+1) - \xi^{(2)}(t)], \tag{31}$$

where $\xi^{(1)}(t) \sim N(0, \sigma^{(1)})$ and $\xi^{(2)}(t)$ has an independent zero-modified normal distribution with $\mathbb{P}(\xi^{(2)}(t) = 0) = 1 - p$, and Gaussian parameters $(\mu^{(2)}, \sigma^{(2)})$. The motivation for (31) is to incorporate idiosyncratic mortality shocks represented by $\xi^{(2)}$, that occur with probability p any given year and have a random magnitude with distribution $N(\mu^{(2)}, \sigma^{(2)})$. Such shocks, representing natural or geopolitical catastrophes, are temporary and last just a single period, hence subtraction of the last term $-\xi^{(2)}(t)$ in (31). Due to this term, it would appear that the model has a two-dimensional state space $\{\kappa^{(2)}(t), \xi^{(2)}(t)\}$. However, we note that it is sufficient to generate scenarios starting with $\kappa^{(2)}(T)$ and assuming $\xi^{(2)}(T) = 0$ (no shock in year T). Then after estimating $f(\kappa) = \mathbb{E}[F(T, \kappa^{(2)}(\cdot))|\kappa^{(2)}(T) = \kappa, \xi^{(2)}(T) = 0]$, one easily obtains in case of year-T shocks $\mathbb{E}[F(T, \kappa^{(2)}(\cdot)|\kappa^{(2)}(T) = \kappa, \xi^{(2)}(T) = \xi] = f(\kappa - \xi)$, reducing to the prediction of "unshocked" values.

The presence of idiosyncratic shocks in $m(t, x)$ renders the corresponding survival probability analytically intractable. However, the linear dynamics of $\kappa^{(2)}$ in (31) allows to obtain the following deterministic estimator for future mortality rates.

Lemma 1. *Let $Z(s) = \{\kappa^{(2)}(s), \xi^{(2)}(s)\}$. Under the Chen-Cox model, the following holds:*

$$\mathbb{E}[\kappa^{(2)}(t) \mid Z(s)] = \kappa^{(2)}(s) + (t - s)\mu^{(1)} + \mu^{(2)}p - \xi^{(2)}(s), \qquad 0 \le s \le t < \infty. \tag{32}$$

The proof can be found in Appendix B. Substituting (32) into (30) yields the following estimator for $\mathbb{E}[m(T + s, x) \mid \kappa^{(2)}(T), \xi^{(2)}(T)]$:

$$\hat{m}(T + s, x) \doteq \exp\left(\beta^{(1)}(x) + \beta^{(2)}(x)\left(\kappa^{(2)}(T) + s\mu^{(1)} + \mu^{(2)}p - \xi^{(2)}(T)\right)\right). \tag{33}$$

4.1. Results

We follow Chen and Cox [14] in using US mortality data obtained from the National Center for Health Statistics (NCHS)[2]. This dataset contains yearly age specific death rates for overall US

[2]Source: `http://www.cdc.gov/nchs/nvss/mortality_tables.htm`

population over 1900–2003. Fitting yields the random-walk parameters $\mu^{(1)} = -0.2173, \sigma^{(1)} = 0.3733$ in (31), as well as the estimated probability of shock as $p = 0.0436$, with jump distribution $(\mu^{(2)}, \sigma^{(2)}) = (0.8393, 1.4316)$. As expected, $\mu^{(2)} \gg 0$ is large and positive, so shocks correspond to large temporary increases in mortality. The goal is to analyze and compare the ability of kriging models and analytic estimates to predict $T = 10$-year deferred annuity values for unisex $x = 65$ year olds. Payments are cut-off at age $\bar{x} = 94$. We use a discount rate of $r = 4\%$.

	$N_{tr} = 125$		$N_{tr} = 512$		$N_{tr} = 1000$	
Type	Bias	$\sqrt{\text{IMSE}}$	Bias	$\sqrt{\text{IMSE}}$	Bias	$\sqrt{\text{IMSE}}$
Analytic	1.668e-03	2.148e-03	1.668e-03	2.148e-03	1.668e-03	2.148e-03
Ord. Kriging	5.145e-03	5.923e-03	1.582e-04	1.975e-03	-1.999e-04	1.634e-03
Univ. Kriging	5.832e-03	6.059e-03	4.816e-04	1.045e-03	-1.243e-05	7.428e-04

Table 1: Comparing estimators for life annuity value under the Chen-Cox model for different size of experimental design. The design $\mathcal{D}$ is constructed with $N_{tr} = N_{tr,1}^{2/3} \cdot N_{tr,2}^{1/3}$. The reported values are evaluated from a Monte Carlo benchmark, using (28) and (29). Analytic estimate is based on (33); universal kriging model uses first-order linear basis functions.

W fit emulators with budgets $N_{tr} \in \{125, 512, 1000\}$. The respective training designs $\mathcal{D}$ are deterministic and uniformly spaced across an appropriately chosen interval $D = [\underline{\kappa}, \bar{\kappa}]$; a fixed design minimizes Monte Carlo noise in fitting $\hat{f}(\cdot)$. Because $Z \equiv \kappa^{(2)}$ is just one-dimensional, a relatively small training budget is used. For the emulators, we fit both an ordinary kriging (OK) model with constant trend $\mu(\kappa) = \beta_0$, and first-order linear universal kriging (UK) model with $\mu(\kappa) = \beta_0 + \beta_1 \kappa$. For evaluation, we fix a testing set containing $N_{out} = 50$ values of $Z(T)$, each with a Monte Carlo benchmark containing $N_{in} = 10^5$ simulations. Due to the very small MSE's involved, a very high-fidelity benchmark was needed (in order to isolate the MSE of the emulator from the MSE of the benchmark), leading to a very large N_{in}. To be computationally feasible, we picked a small testing set. To make sure that $\mathcal{D}^{test}$ accurately represents the distribution of $Z(T)$ its locations were picked as the empirical $1\%, 3\%, \ldots, 99\%$ percentiles of a large sample of $Z(T)$. The mortality shocks associated with these percentiles were used in the comparison process.

Table 1 and Figure 1 summarize the results. We observe that there is quite a wide spread in potential future annuity prices, with differences of more than 10% (or \$1 in annuity NPV) depending on realized $Z(T)$. This confirms the significant level of longevity risk. As shown in the Figure 1, there is a nearly linear relationship for $z \mapsto a(z, T, x)$, which is perhaps surprising given the above range of forecasts. This strong linear trend in the response partly explains the advantage of the UK model over OK. The Figure also reflects the effect of training set size and distribution: the $N_{tr} = 512$ model performs significantly better than its $N_{tr} = 125$ counterpart. We see that all methods perform well, with IMSE's on the order of 1e-03. Even though the computed biases are rather small, we remark that since pension portfolios have very large face values, the corresponding approximation errors could be financially meaningful. For example, for a modest pension fund with an obligation of \$100mm, a bias of 1e-03 implies inaccuracy of \$100k.

The right panel of Figure 1 provides a zoomed-in visualization of the estimators' bias relative to $\hat{a}^{MC}$. As expected, the analytic estimator based on Lemma 1 overestimates the true annuity value for all $\kappa^{(2)}(T)$. For $N_{tr} = 125$, the kriging emulator clearly has a larger MSE, and in this case typically under-estimates $a(\kappa^{(2)}(T))$. For $N_{tr} = 512$ we observe the statistical learning taking place, as the kriging model now has an excellent fit in the middle of the plot and essentially zero

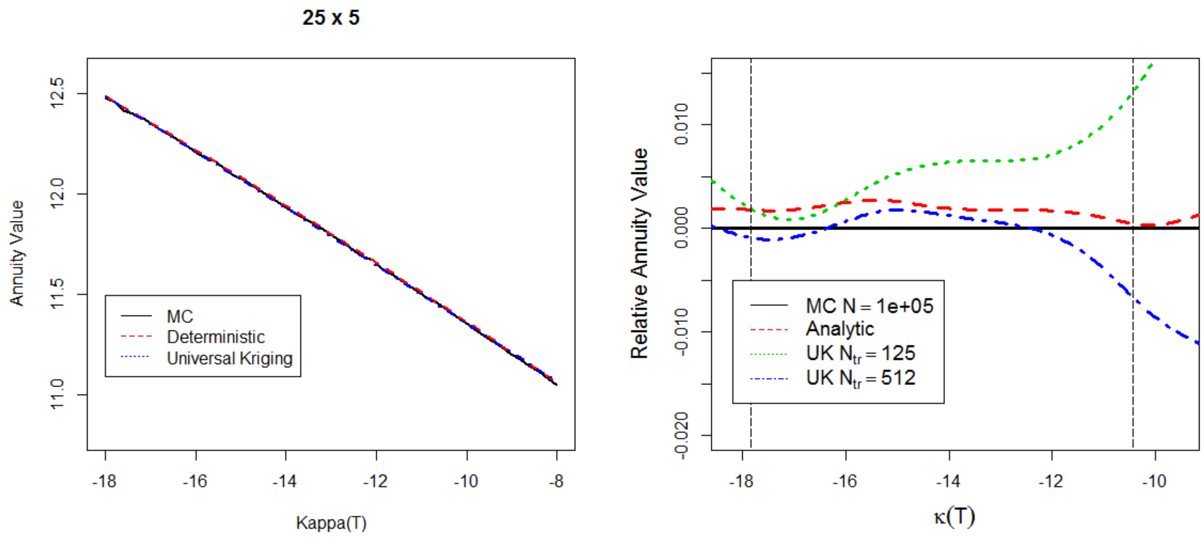

Figure 1: Annuity emulators in the Chen-Cox model. Left: three estimators (MC, UK w/$N_{tr} = 125$ and analytic) of annuity value $a(\kappa^{(2)}(T))$ vs. $\kappa^{(2)}(T)$. The training design (indicated by the vertical dashed lines) is $\mathcal{D} = \{\kappa^{(2)}(T) \in (-17.5, -10)\}$ with $N_{tr,1} = 25, N_{tr,2} = 5$. Right: relative annuity values vis-a-vis the Monte Carlo benchmark $\hat{a}^{MC}$ obtained with $N_{in} = 10^5$.

bias averaging over potential values of $\kappa^{(2)}(T)$. The effect of larger training budget is confirmed in Table 1, with IMSE's all decreasing towards zero as N_{tr} increases.

The above analysis demonstrates that in some settings, the shape $z \mapsto a(z, T, x)$ is sufficiently simple that little modeling is required, and analytic estimators perform well (as do statistical emulators). However, we stress that there is no easy way to tell a priori that the analytic estimator would be adequate, and in any case a sufficiently large training set size will guarantee a better predictive power for the kriging models.

The one dimensional case also provides a visual representation of the effect of grid design, illustrated in Figure 2. The figure showcases two features of emulators: (i) dependence between local accuracy as measured by $s^2(z)$ and grid *size* N_{tr}; and (ii) dependence between $s^2(z)$ and grid *shape*. First, larger training sets improve accuracy (with a general relationship of $\mathcal{O}(N_{tr}^{-1/2})$ like in plain Monte Carlo). This can be seen in Figure 2 where kriging standard deviation $s(z)$ is consistently lower for $N_{tr}^{(B)} = 1000$ compared to $N_{tr}^{(A)} = 125$. One implication is that as $N_{tr} \to \infty$, we would have $s^2(z) \to 0$, i.e. $f(\cdot)$ would be learned with complete precision, a property known as global consistency of the emulator. Second, $s^2(z)$ is affected by the shape of $\mathcal{D}$ in the sense that higher local density of training points lowers the local posterior variance. This is intuitive if viewing $\hat{f}$ as an interpolator or kernel regressor – the denser the training set around z, the better we are able to infer $f(z)$. Consequently, the empirical grid $\mathcal{D}^{(C)}$ that is concentrated around the mode of $Z(T)$, offers better accuracy in that neighborhood (around $\kappa^{(2)}(T) \simeq -14$ in Figure 2) compared to the uniform $\mathcal{D}^{(B)}$, but lower accuracy towards the edges, where $\mathcal{D}^{(C)}$ becomes sparser. For all designs, posterior uncertainly deteriorates markedly as we migrate outside of the training set (e.g. $\kappa^{(2)}(T) > -11$ in the Figure).

The $s(z)$ values shown in Figure 2 also provide an approximation of emulator IMSE. For

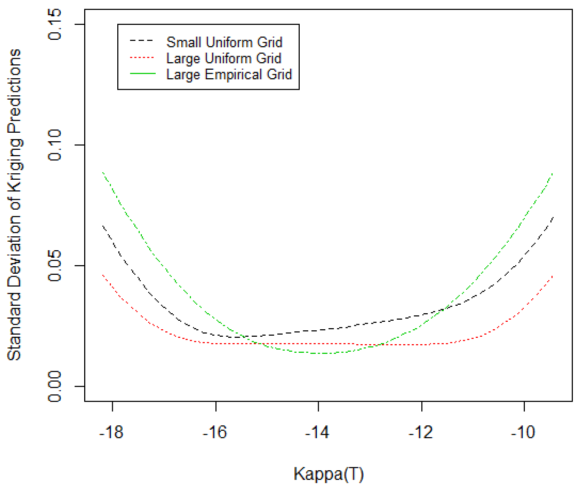

Figure 2: Effect of training design $\mathcal{D}$ on the emulator accuracy in the Chen-Cox model. We show kriging standard deviation $s(z)$ for the universal kriging model with three different designs: $\mathcal{D}^{(A)}$ (small uniform), $\mathcal{D}^{(B)}$ (large uniform) and $\mathcal{D}^{(C)}$ (large empirical). The deterministic designs $\mathcal{D}^{(A)}, \mathcal{D}^{(B)}$ contain uniformly spaced values of $\kappa^{(2)}(T) \in (-17.5, 10)$, of size $N_{tr}^{(A)} = 125$ and $N_{tr}^{(B)} = 1000$ respectively. $\mathcal{D}^{(C)}$ is an empirical design of size $N_{tr}^{(C)} = 1000$ generated using the density of $\kappa^{(2)}(T)|\kappa^{(2)}(0)$.

example, averaging the kriging standard deviation $s(z)$ over the testing set using the UK model with $N_{tr} = 1000$ yields $s_{Ave} = \sqrt{\{\frac{1}{N_{out}} \sum_n s^2(z^{(n)})\}} = 7.159e - 03$, while in Table 1 the corresponding reported IMSE was $\widehat{IMSE} = 7.428e - 04$. Reasons for the mismatch include the residual MSE in the Monte Carlo estimate $\hat{a}^{MC}$ and model mis-specification of the UK model, which would bias the self-assessed accuracy. Moreover, the strong correlation between $\hat{a}(z)$ across different testing locations $z^{(n)}$ implies that $\widehat{IMSE}$ has a large standard error. Nevertheless, s_{Ave} is a highly useful metric that allows to quantify the relative accuracy of different emulators in the absence of any gold-standard benchmarks.

5. Case Study: Hedging an Index-Based Fund in a Two-Population Model

There has been a lot of recent discussion regarding index-based longevity funds. Information on the death rates of the general public is widely available, and a market fund that uses the respective death rats as its price index offers a standardized way to measure population longevity. In particular, it allows for securitization of longevity swaps that can be used by pension funds to hedge their longevity risk exposure. If the pension fund could buy as many units of the swap as it has to pay out to its annuitants, it would result in a situation where the amount paid is nearly equal to the amount received from the swap. The quality of such as hedge is driven by the basis risk between the indexed population and the annuitant pool, that is typically a subset of the index. Consequently, it is necessary to create a model to capture the link between the index and the insured sub-population.

Remark. From a different angle, some longevity products explicitly integrate mortality experience in several regions, for example across different countries (UK, Germany, Netherlands) or across different constituencies (England vis-a-vis Great Britain). Lin et al. [30] states that most mortality data reported by official agencies calculate a weighted average mortality index of different underlying populations. They also investigate the modeling aspect of such multi-population indices.

To fix ideas, we call the index population Pool 1, and the annuitants Pool 2. Consider now an individual from Pool 2 who will be aged x at date T when she begin to receive her life annuity. The corresponding time-T liability to the pension fund is denoted $a_2(Z(T), T, x)$. If the pension fund enters into a swap based on the index, she might purchase π index-fund annuities for age x, with net present value of $\pi a_1(Z(T), T, x)$, at T. For now we ignore what would be a fixed premium. The overall hedge portfolio is then $\Delta(Z(T), T, x) \doteq \pi a_1(Z(T), T, x) - a_2(Z(T), T, x)$. Several risk measures can be used to determine hedge effectiveness. Some examples include variance, or tail risk measures such as value-at-risk (VaR) or expected shortfall (TVaR). Recent work in this direction includes Coughlan et al. [15] who used a bootstrapping and extrapolation method to analyze hedge effectiveness, and Cairns et al. [13] whose setup we follow below.

Unsurprisingly, the correlation structure for mortality across populations is complex. One notable recent contribution is by Cairns et al. [12, 13] who considered a hedging problem between an index pool $k = 1$ and insured sub-pool $k = 2$. Specifically, the two populations are the England & Wales (E&W) general population, which represents the index mortality rate (Pool 1), and the Continuous Mortality Investigation (CMI) population, which are mortality rates gathered from United Kingdom insured populations, serving the role of those receiving pension payments (Pool 2). To model the dependence between the two pools, Cairns et al. [12] proposed a cointegrated two-population Bayesian model based on the Lee-Carter framework. To wit, the mortality rates $m_k(t, x)$ behave similar to (11),

$$\log m_k(t, x) = \beta_k^{(1)}(x) + n_a^{-1}\kappa_k^{(2)}(t) + \gamma_k^{(3)}(t - x), \quad k = 1, 2$$

with stochastic dynamics of the period effect $\kappa_1^{(1)}$ given by

$$\kappa_1^{(2)}(t) = \kappa_1^{(2)}(t-1) + \mu_1 + \sigma_1\epsilon_1(t), \quad \epsilon_1(t) \overset{\text{i.i.d}}{\sim} N(0, 1). \tag{34}$$

In turn, the mortality of the larger population influences the period effect of the smaller (insured) population, with dynamics for $\kappa_2^{(2)}$ co-integrated with $\kappa_1^{(1)}$. Namely, their difference $S(t) \doteq \kappa_1^{(2)}(t) - \kappa_2^{(2)}(t)$ forms an $AR(1)$ process

$$S(t) = \mu_2 + \phi(S(t-1) - \mu_2) + \sigma_2\epsilon_2(t-1) + c\epsilon_1(t-1), \quad \epsilon_2(t) \overset{\text{iid}}{\sim} N(0, 1), \tag{35}$$

with $\epsilon_1(\cdot)$ being independent of $\epsilon_2(\cdot)$, and $c = \sigma_1 - \rho\sigma_2$ for the covariance, where $\rho = \mathrm{Corr}(\kappa_1^{(2)}(t), \kappa_2^{(2)}(t))$. In both models cohort effects $\gamma_k^{(3)}$ are independent $AR(2)$ processes. Here, ϕ is the mean-reversion rate. Since (35) models the difference $\kappa_1^{(2)}(t) - \kappa_2^{(2)}(t)$, ϕ reflects the rapidity of how $S(t)$ returns to the constant μ_2, which is assumed to be the constant difference between the two populations.

5.1. Analytic Approximations

Cairns et al. [13] used the fact that $\mathbb{E}[\kappa_1^{(2)}(T+t) \mid \kappa_1^{(2)}(T)] = \kappa_1^{(2)}(T) + \mu_1 t$ to introduce the median-mortality approximation

$$\hat{m}_1^{A1}(T+t, x+t) = \exp\left(\beta_1^{(1)}(x+t) + \frac{1}{n_a}(\kappa_1^{(2)}(T) + \mu_1 t) + \frac{1}{n_a}\gamma_1^{(3)}(T-x)\right). \tag{36}$$

Since $S(t)$ is mean reverting, it is also suggested to use the approximation for the CMI population of

$$\hat{m}_2^{A1}(T+t,x+t) = \exp\left(\beta_2^{(1)}(x+t) + \frac{1}{n_a}(\kappa_2^{(2)}(T)+\mu_1 t) + \frac{1}{n_a}\gamma_2^{(3)}(T-x)\right), \tag{37}$$

i.e. the same drift as the general population but different initial value.

We introduce a different, more accurate approximation based on the following lemma.

Lemma 2. *We have*

$$\mathbb{E}\left[\kappa_2^{(2)}(T+t)|Z(t)\right] = \kappa_1^{(2)}(T) + \mu_1 t - \mu_2(1-\phi^t) - \phi^t(\kappa_1^{(2)}(T) - \kappa_2^{(2)}(T)). \tag{38}$$

The proof can be found in Appendix B. Denote $\mathbb{E}[\kappa_2^{(2)}(T+t)|Z(t)] \doteq \xi(t,T)$. Lemma 2 suggests an alternative analytic estimator for $m_2(T+s,x)$ as

$$\hat{m}_2^{A2}(T+s,x+s) \doteq \exp\left(\beta_2^{(1)}(x+s) + \xi(t,T) + \frac{1}{n_a}\gamma_2(T-x)\right). \tag{39}$$

Denote $a_1(Z(T))$ and $a_2(Z(T))$ as the net present value at T (conditional on $Z(T)$) of a life annuity for the E&W and CMI populations respectively as defined in (9). In what follows, *Analytic 1* will refer to use of (36) and (37) in estimating survival probabilities (10) for each population (and hence a_1 and a_2), while *Analytic 2* refers to the use of (36) and (39). Notation for deferred annuity values under the two analytic approaches will be $\hat{a}_k^{A1}(z)$ and $\hat{a}_k^{A2}(z), k=1,2$.

5.2. Model Fitting

The parameters $\beta_1^{(1)}(x)$, $\beta_2^{(1)}(x)$, and past trajectories $\kappa_k^{(2)}(t)$, $\gamma_k^{(3)}(t-x)$, for $k=1,2$ were estimated from the male E&W and CMI populations respectively, and the time and age ranging from calendar years 1961 to 2005 (with 2005 treated as $t=0$), and x from 50 to 89. The processes $(\kappa_1^{(2)}(t))$ and $(S(t))$ were fit as random walk with drift and $AR(1)$ respectively, introducing additional parameter estimates for $\mu_1, \sigma_1, \mu_2, \phi, \sigma_2$ and c. We find $\mu_1 = -0.5504, \mu_2 = 0.6105, \sigma_1 = 1.278, \sigma_2 = 0.568, \phi = 0.9407, c = 0.262$, so that the CMI population tends to have higher mortality, with a co-integration of about 94%.

Using the PPC approach of [13], we treat the age-effect parameters as fixed, and refit the $ARIMA$ models at period T for each simulation. That is, the $\beta_k^{(1)}$ are fixed throughout for $k=1,2$ and each of μ_1, σ_2, μ_2, ϕ, σ_2, and c are re-estimated. In principle, this makes the re-estimated parameters part of the state variable $Z(T)$. A few preliminary runs indicate that the variance parameters σ_1, σ_2 and c have little significant effect on annuity values, while μ_1, μ_2 and ϕ do. Since μ_1 is in one-to-one correspondence with $\kappa_1^{(2)}(T)$, our time T state process is finally characterized as

$$Z(T) = \{\kappa_1^{(2)}(T), \kappa_2^{(2)}(T), \mu_2, \phi\}.$$

Heuristically, this is a reasonable choice: each element of $Z(T)$ has a direct effect on the time T mortality rates or their trends, while the variance terms simply add variability.

Several stochastic mortality models have `R` code available[3] for model fitting. We use the code to fit the two-population model parameters, yielding the inferred past trajectories for the age, period,

[3]LifeMetrics Open Source R code for Stochastic Mortality Modelling; see `http://www.macs.hw.ac.uk/~andrewc/lifemetrics/` for details

and cohort effects. In a separate step, the estimated period and cohort effects are modeled as individual $ARIMA$ models.

For the remainder of this section we assume the starting age of the annuitant is $x = 65$ with a fixed interest rate of $r = 0.04$ and a $T = 10$ year deferral period. Generally the choice of hedge ratio π is chosen systematically, for example through minimizing variance. In this paper we assume the neutral value of $\pi = 1$ in order to not favor one estimation type over another. Hence the value of the hedge portfolio is $\Delta(Z(T)) = a_1(Z(T)) - a_2(Z(T))$.

As discussed in Section 3.3, determining the training set design depends on the problem at hand. In our particular example with a 4-dim $Z(T)$, we aim to give an accurate result of the expectation of the hedge portfolio $\Delta(T)$, so we use an empirical design, as suggested in Section 3.3. This also holds the advantage of capturing the correlation between $\kappa^{(1)}$ and $\kappa^{(2)}$ which is important in this co-integrated model. To compare the effect of budget size, we choose two different budgets, $N_{tr} = 1000$ and $N_{tr} = 8000$. Following the framework in Section 3.3, N_{tr} is allocated into $N_{tr,1} = N_{tr}^{2/3}, N_{tr,2} = N_{tr}^{1/3}$, so that we have $N_{tr,1} = 100$ (resp. $N_{tr,1} = 400$) training points with Monte Carlo simulations containing $N_{tr,2} = 10$ (resp. $N_{tr,2} = 20$) batched simulations for each design point.

Different surrogate models are chosen than in Section 4; this time around a multi-dimensional state process suggests the use of a TPS model from Section 3.1. We forego the OK model, but maintain use of the 1st-order linear UK model, and also implement a simple kriging (SK) model with $\mu(z) = \hat{a}_1^{A2}(z) - \hat{a}_2^{A2}(z)$. This combines advantages from both the analytic and UK approach, giving us an already accurate estimate for the trend, while nonparametrically modeling the residuals. For these reasons, a SK emulator should outperform both the analytic estimators and the UK model. We utilize another advantage of the surrogate models and fit them directly to the hedge portfolio values $\Delta(Z(T))$ rather than individually modeling annuity values $a_k(Z(T))$ and then taking difference of two approximations.

The Monte Carlo benchmark yields an average portfolio value of 0.1995 with a standard deviation of 0.1067. This suggests that a one-to-one purchase of index annuity is not the optimal hedge unit under this population model. In an actual application, one would analyze further to determine a different choice for π; for example Cairns et al. [13] chooses π to minimize portfolio variance.

5.3. Results

	$N_{tr} = 1000$		$N_{tr} = 8000$	
Type	Bias	$\sqrt{\text{IMSE}}$	Bias	$\sqrt{\text{IMSE}}$
Analytic A1 from (37)	-2.101e-02	3.460e-02	-2.101e-02	3.460e-02
Analytic A2 from (39)	4.480e-03	5.321e-03	4.480e-03	5.321e-03
Thin Plate Spline	2.577e-03	1.304e-02	5.803e-04	5.095e-03
Universal Kriging	4.363e-04	1.856e-02	1.857e-03	1.289e-02
Simple Kriging	-1.334e-03	3.280e-03	9.390e-04	3.043e-03

Table 2: Performance of analytic estimates and surrogate models for hedge portfolio values in the two-population model case study. Numbers reported are based on $N_{out} = 1000$ simulations of $Z(T)$ with a Monte Carlo benchmark. N_{tr} is allocated into $N_{tr,1} = N_{tr}^{2/3}$ training points and $N_{tr,2} = N_{tr}^{1/3}$ Monte Carlo batches per training point. Simple kriging model uses A2 estimator as trend.

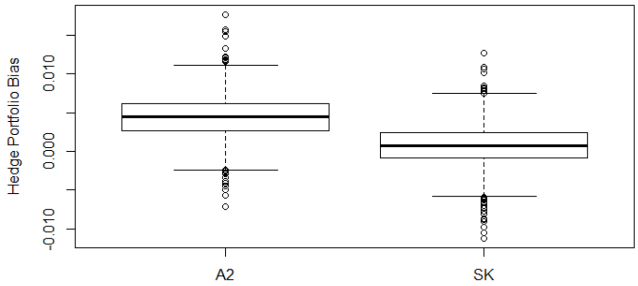

Figure 3: Boxplots of hedge portfolio value bias for $N_{tr} = 8000$ for analytic A2 and simple kriging approaches. To construct the boxplot, we computed for each of 1000 simulated values of $Z(T)$, the difference between the respective estimate and the Monte Carlo benchmark.

We choose $N_{out} = 1000$ simulations of $Z(T)$ and predict hedge portfolio values $\Delta(Z(T))$ with the surrogate models, as well as via the deterministic estimates. Table 2 shows the results. As expected, the Analytic A2 estimator outperforms Analytic A1 since it is catered directly to the two-population model. Relative to A1, our improved estimator cuts bias by nearly 80%. As for the surrogates, when $N_{tr} = 1000$, each of the TPS and UK models only slightly underperform the analytic estimate A2, while the SK model does significantly better. For $N_{tr} = 8000$, both TPS and SK are better than A2.

Figure 3 summarizes the empirical distribution of the bias of the A2 and SK estimators given simulations of $Z(T)$. We can see that both approaches have similar variability, while SK has a much lower bias. The UK and TPS estimators have similar distributions with slightly larger bias than SK.

There are a few comments to be made in regards to these results. First of all, there is no way to tell a priori that a deterministic estimate will perform well. For example each surrogate model completely outclasses A1, while TPS and UK perform only marginally better than A2. Possibly, even better (or worse) analytic estimators can be derived. Additionally, the deterministic estimators are for annuity values themselves and not for the portfolio difference $\Delta(T)$. A lower bias for a $\Delta(T)$ could simply be a consequence of the bias of each annuity $a_k(Z(T))$ being reduced during subtraction.

6. Case Study: Predicting Annuity Values under the CBD Framework

6.1. Model Fitting

Our third case study utilizes another popular class of mortality models, the CBD [9] models which directly work with the survival probabilities P. To wit, we model the 1-period survival probability

$$P(Z(T); T, 1, x) = \frac{1}{1 + \exp\left(\kappa^{(1)}(T) + (x - x_{Ave})\kappa^{(2)}(T)\right)}, \tag{40}$$

where $x_{Ave} = n_a^{-1} \sum_i x_i$, and $\kappa^{(1)}, \kappa^{(2)}$ follow $ARIMA$ models, which according to Cairns et al. [10] provide a good fit for period effects. Multi-period survival probabilities are obtained as products of (40).

We fit (40) to the CMI population, considering a full range of $ARIMA(p, d, q)$ models with $p, q = 0, 1, 2, 3, 4$ and $d = 0, 1, 2$, using `auto.arima` in `R` from the package "forecast" [23]. The optimal configuration for this population is for $\kappa^{(1)}$ to follow $ARIMA(0, 1, 3)$ with drift and $\kappa^{(2)}$ to follow $ARIMA(1, 1, 2)$:

$$\kappa^{(1)}(t) = \kappa^{(1)}(t-1) + \mu + \epsilon^{(1)}(t) + \sum_{q=1}^{3} \theta^{(q,1)} \epsilon^{(1)}(t-q), \tag{41}$$

$$\kappa^{(2)}(t) = (1+\phi)\kappa^{(2)}(t-1) - \phi\kappa^{(2)}(t-2) + \epsilon^{(2)}(t) + \sum_{q=1}^{2} \theta^{(q,2)} \epsilon^{(2)}(t-q). \tag{42}$$

The estimated $ARIMA$ parameters are $\mu = -0.0195, \phi = 0.9206, \theta^{(1,1)} = -0.5516, \theta^{(2,1)} = 0.1736, \theta^{(3,1)} = 0.5169, \theta^{(2,1)} = -1.4664, \theta^{(2,2)} = 0.6167$. The $\theta^{(q,k)}, k = 1, 2$ describe how past errors echo into future values of $\kappa^{(k)}$. For example, the large negative value of $\theta^{(2,1)}$ means that the noise generated in $\kappa^{(1)}(s)$ will be amplified, made negative, and added to the future $\kappa^{(1)}(s+2)$. The above equations imply that the state has three components, $Z(T) = \{\kappa^{(1)}(T), \kappa^{(2)}(T), \kappa^{(2)}(T-1)\}$.

As in the previous case studies, we develop a deterministic estimate for survival probabilities. Denote by $\xi^{(k)}(t, s) \doteq \mathbb{E}[\kappa^{(k)}(t) \mid Z(s)]$ for $k = 1, 2$. The expressions for $\xi^{(k)}$ are as follows.

Lemma 3. *The following hold for $t > s$*

$$\xi^{(1)}(t, s) = \kappa^{(1)}(s) + \mu(t-s); \tag{43}$$

$$\xi^{(2)}(t, s) = \phi^{t+1-s} \left(\frac{\kappa^{(2)}(s) - \kappa^{(2)}(s-1)}{\phi - 1} \right) + \left(\frac{\phi\kappa^{(2)}(s-1) - \kappa^{(2)}(s)}{\phi - 1} \right). \tag{44}$$

The proof can be found in Appendix B.3. Based on Lemma 3, and substituting expected values of $\kappa^{(k)}(s)$ into (40) we obtain a deterministic estimate of the u-year survival probability as the product

$$\hat{P}^{det}(Z(s), t, u, x) = \prod_{j=0}^{u-1} \frac{1}{1 + \exp\left(\xi^{(1)}(t+j, s) + (x + j - x_{Ave})\xi^{(2)}(t+j, s)\right)}.$$

Through equation (9), this yields the estimate for the $T-$year deferred annuity given $Z(T)$:

$$\hat{a}^{det}(Z(T); T, x) = \sum_{s=1}^{\bar{x}-x} e^{-rs} \hat{P}(Z(T), T, s, x), \tag{45}$$

where the cutoff age is $\bar{x} = 89$.

We proceed to value life annuities in the above model. In contrast to the first two case studies, we extend the deferral period to twenty years. An additional ten years of evolution imbues significant uncertainty into the mortality state $Z(T)$. We use an empirical training design $\mathcal{D}$ for this case study for two reasons; one being that the correlation structure is problematic with any organized grid. From (42), we see that $\kappa^{(2)}(20)$ and $\kappa^{(2)}(19)$ should be strongly correlated, while both $\kappa^{(2)}(19)$ and $\kappa^{(2)}(20)$ are independent of $\kappa^{(1)}(20)$. Secondly, the long deferral period causes significant variation in the distribution of $Z(20)$, and with expectation in mind, we desire the accurate capturing of the density of $Z(20)$ that the empirical grid will provide. The algorithms discussed in Sections 3.3 and 3.4 are used to generate the design and fit the surrogate models. As in Section 4, we choose an ordinary kriging and 1st-order linear universal kriging models, and also fit a thin plate spline model as used in Section 5.

6.2. Results

Type	$N_{tr} = 1000$ Bias	$\sqrt{\text{IMSE}}$	$N_{tr} = 8000$ Bias	$\sqrt{\text{IMSE}}$
Analytic	-4.560e-01	5.257e-01	-4.560e-01	5.257e-01
Spline	-2.358e-02	6.719e-02	4.195e-03	5.436e-02
Ord. Kriging	3.669e-03	9.785e-02	9.734e-03	7.743e-02
Univ. Kriging	-1.785e-03	5.844e-02	5.635e-03	4.355e-02

Table 3: Performance of analytic estimates and surrogate models for 20-year deferred annuity values under the CBD framework. Numbers reported are based on $N_{out} = 1000$ draws of $Z(20)$. N_{tr} is allocated into $N_{tr,1} = N_{tr}^{2/3}$ training points and $N_{tr,2} = N_{tr}^{1/3}$ Monte Carlo batches per training point. Analytic estimate refers to (45), and Spline to thin plate spline (TPS) model. Universal kriging model uses linear basis functions.

In contrast to the results in Sections 4.1 and 5.3, Table 3 shows that the analytic estimator (45) crumbles under this volatile model and long deferral period. On the other hand, both kriging models produce reasonable results even with $N_{tr} = 1000$. We can also observe a diminished effect of increasing the training set size, due to the increased model variance.

These results reflect the comments made in the previous sections: the analytic estimate is a parametric guess as to what may provide an accurate result, and that guess is not always correct. Our analytic choice in this case study was derived along identical lines as to the analytic estimates in the other case studies, yet performs substantially worse. In comparison, the statistical learning frameworks provide a reliable estimator even in a volatile model with a three-dimensional state process and long deferral period.

7. Conclusion

The three case studies above showcase the flexibility and admirable performance of the surrogate models across a range of various longevity risk dynamics. Compared to the consistent accuracy of the statistical emulators, the quality of the deterministic projections was widely varying. Because an analytic derivation is required to produce a deterministic estimator, there are several plausible estimators available. In Section 5 we derived two different estimators, both of which were viable, but one underperformed. Similarly, in Section 6 the derived deterministic projection was also inaccurate. Overall, these examples show that our models can outperform deterministic projections and provide minimally-biased estimates.

Relative to the analytic estimates, the case studies in section 5 required the training set size to be sufficiently large, while in Sections 4-6 a small design of less than 10^3 simulations yielded accurate models. In terms of individual model performance, it was without surprise that the SK model in Section 5 produced the best results. The analytic estimate was already performing adequately, and the SK model used it as its trend component to improve accuracy even more. The downside to this is that a well-performing deterministic estimate was required.

Throughout the paper we have hinted at possibilities for further work. Straightforward extensions include using other mortality models, or emulating other insurance products, e.g. variable life annuities. One can also build more dynamic surrogates that treat initial age x (fixed in our case studies as $x = 65$) or deferral period T as part of the state Z, providing a joint prediction for

$(Z(0), T, x) \mapsto \mathbb{E}[a(Z(T), T, x)]$. Similarly, one could consider more parameter uncertainty which would lead to including additional components in the state Z.

The emulators we obtained for $a(Z(T), T, x)$ offer a high-performance tool for annuity risk management. Indeed, they are based on advanced, previously vetted stochastic mortality models and calibrated to real, reliable, large-scale mortality datasets. Hence, the fitted estimates for annuity values are in essence a best-available forecast that combines state-of-the-art longevity modeling, data calibration and statistical model. As such, (after incorporating age and interest rate as model parameters) they would be of independent interest to actuaries working in longevity space and seeking easy-to-use tools for forecasting net present values of life annuities. The emulator offers a plug-and-play functionality, converting inputted parameters (such as age x, deferral period T and discount rate r) into the annuity value (note that the initial state $Z(0)$ is read off from the calibration procedure). One can imagine building a library of such emulators for different mortality-contingent products available in the marketplace.

Looking more broadly, the emulation approach we propose is very general and can be applied in a variety of actuarial contexts. In particular, in future work we plan to extend it to the microscopic agent-based models of mortality Barrieu et al. [2] which offer a canonical "complex system" representation of population longevity. We believe that emulators could significantly simplify predictions in these types of models by providing a tractable, statistical representation of demographic interactions within a stochastic dynamic population framework. Another class of insurance applications requires functional-regression tools where emulators can again be very effective [20]. A different extension is emulation of risk measures related to $F(T, Z(\cdot))$, such as VaR or TVaR, which require targeted surrogates that focus on a specific region of the input space. A starting point is to combine concept of importance sampling to generate a targeted design $\mathcal{D}$ that e.g. preferentially concentrates on the left tail of F.

References

[1] Bacinello, A., Biffis, E., Millossovich, P., 2010. Regression-based algorithms for life insurance contracts with surrender guarantees. Quantitative Finance 10 (9), 1077–1090.

[2] Barrieu, P., Bensusan, H., El Karoui, N., Hillairet, C., Loisel, S., Ravanelli, C., Salhi, Y., 2012. Understanding, modelling and managing longevity risk: key issues and main challenges. Scandinavian Actuarial Journal 2012 (3), 203–231.

[3] Bauer, D., Benth, F. E., Kiesel, R., 2012. Modeling the forward surface of mortality. SIAM Journal on Financial Mathematics 3 (1), 639–666.

[4] Bauer, D., Reuss, A., Singer, D., 2012. On the calculation of the solvency capital requirement based on nested simulations. ASTIN Bulletin 42 (02), 453–499.

[5] Booth, H., Maindonald, J., Smith, L., 2002. Applying Lee–Carter under conditions of variable mortality decline. Population Studies 56 (3), 325–336.

[6] Boyer, M. M., Stentoft, L., 2013. If we can simulate it, we can insure it: An application to longevity risk management. Insurance: Mathematics and Economics 52 (1), 35–45.

[7] Broadie, M., Du, Y., Moallemi, C. C., 2011. Efficient risk estimation via nested sequential simulation. Management Science 57 (6), 1172–1194.

[8] Brouhns, N., Denuit, M., Vermunt, J. K., 2002. A Poisson log-bilinear regression approach to the construction of projected lifetables. Insurance: Mathematics and Economics 31 (3), 373–393.

[9] Cairns, A. J., Blake, D., Dowd, K., 2006. A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. Journal of Risk and Insurance 73 (4), 687–718.

[10] Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Khalaf-Allah, M., 2011. Mortality density forecasts: An analysis of six stochastic mortality models. Insurance: Mathematics and Economics 48 (3), 355–367.

[11] Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., Balevich, I., 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. North American Actuarial Journal 13 (1), 1–35.

[12] Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Khalaf-Allah, M., 2011. Bayesian stochastic mortality modelling for two populations. ASTIN Bulletin 41 (01), 29–59.

[13] Cairns, A. J., Dowd, K., Blake, D., Coughlan, G. D., 2014. Longevity hedge effectiveness: A decomposition. Quantitative Finance 14 (2), 217–235.

[14] Chen, H., Cox, S. H., 2009. Modeling mortality with jumps: Applications to mortality securitization. Journal of Risk and Insurance 76 (3), 727–751.

[15] Coughlan, G. D., Khalaf-Allah, M., Ye, Y., Kumar, S., Cairns, A. J., Blake, D., Dowd, K., 2011. Longevity hedging 101: A framework for longevity basis risk analysis and hedge effectiveness. North American Actuarial Journal 15 (2), 150–176.

[16] Czado, C., Delwarde, A., Denuit, M., 2005. Bayesian Poisson log-bilinear mortality projections. Insurance: Mathematics and Economics 36 (3), 260–284.

[17] Debonneuil, E., 2010. A simple model of mortality trends aiming at universality: Lee Carter + Cohort. Tech. rep., arXiv:1003.1802.

[18] Delwarde, A., Denuit, M., Eilers, P., 2007. Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting a penalized log-likelihood approach. Statistical Modelling 7 (1), 29–48.

[19] Fushimi, T., Kogure, A., 2014. A Bayesian approach to longevity derivative pricing under stochastic interest rates with a two-factor Lee-Carter model. Tech. rep., ARIA 2014 Annual Meeting.
URL http://www.aria.org/Annual_Meeting/2014/2014_Accepted_Papers/4C/FushimiandKogure.pdf

[20] Gan, G., Lin, X. S., 2015. Valuation of large variable annuity portfolios under nested simulation: A functional data approach. Insurance: Mathematics and Economics 62, 138–150.

[21] Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning, 2nd Edition. Springer Series in Statistics. Springer.

[22] Helbert, C., Dupuy, D., Carraro, L., 2009. Assessment of uncertainty in computer experiments from Universal to Bayesian Kriging. Applied Stochastic Models in Business and Industry 25 (2), 99–113.

[23] Hyndman, R. J., 2015. forecast: Forecasting functions for time series and linear models. R package version 6.1.
URL http://github.com/robjhyndman/forecast

[24] Hyndman, R. J., Ullah, M. S., 2007. Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics & Data Analysis 51 (10), 4942–4956.

[25] Jalen, L., Mamon, R., 2009. Valuation of contingent claims with mortality and interest rate risks. Mathematical and Computer Modelling 49 (9), 1893–1904.

[26] Kleijnen, J. P., 2015. Design and analysis of simulation experiments, 2nd Edition. Vol. 111. Springer Science & Business Media.

[27] Lee, R., Miller, T., 2001. Evaluating the performance of the Lee–Carter method for forecasting mortality. Demography 38 (4), 537–549.

[28] Lee, R. D., Carter, L. R., 1992. Modeling and forecasting US mortality. Journal of the American Statistical Association 87 (419), 659–671.

[29] Li, J.-H., Hardy, M., Tan, K., 2009. Uncertainty in model forecasting: An extension to the classic Lee-Carter approach. ASTIN Bulletin 39, 137–164.

[30] Lin, Y., Liu, S., Yu, J., 2013. Pricing mortality securities with correlated mortality indexes. Journal of Risk and Insurance 80 (4), 921–948.

[31] Nychka, D., Furrer, R., Sain, S., 2015. fields: Tools for Spatial Data. R package version 8.2-1.
URL http://CRAN.R-project.org/package=fields

[32] Plat, R., 2009. On stochastic mortality modeling. Insurance: Mathematics and Economics 45 (3), 393–404.

[33] Rasmussen, C. E., Williams, C. K. I., 2006. Gaussian Processes for Machine Learning. The MIT Press.

[34] Renshaw, A. E., Haberman, S., 2006. A cohort-based extension to the Lee–Carter model for mortality reduction factors. Insurance: Mathematics and Economics 38 (3), 556–570.

[35] Roustant, O., Ginsbourger, D., Deville, Y., 2012. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of Statistical Software 51 (1), 1–55.
URL http://www.jstatsoft.org/v51/i01/

[36] Santner, T. J., Williams, B. J., Notz, W. I., 2003. The Design and Analysis of Computer Experiments. Springer-Verlag, New York, NY.

[37] Sobol, I. M., 1998. On quasi-Monte Carlo integrations. Mathematics and Computers in Simulation 47 (2),

103–112.
[38] Wyss, G. D., Jorgensen, K. H., 1998. A user's guide to LHS: Sandia's Latin hypercube sampling software. Tech. rep., SAND98-0210, Sandia National Laboratories, Albuquerque, NM.

Appendix A. Lee Carter & CBD Stochastic Mortality Models

In this section we give a brief summary of existing stochastic mortality models. We use the notation of Cairns et al. [10] who provided a comprehensive comparison of several mortality models using CMI data.

The APC Lee-Carter model (introduced by Renshaw and Haberman [34]) models the log mortality rate as

$$\log m(t,x) = \beta^{(1)}(x) + \beta^{(2)}(x)\kappa^{(2)}(t) + \beta^{(3)}(x)\gamma^{(3)}(t-x). \tag{M2}$$

One can interpret $\beta^{(1)}(x)$, $\kappa^{(2)}(t)$ and $\gamma^{(3)}$ as the age, period and cohort effects, respectively. The original model proposed by Lee and Carter [28] is a special case where $\gamma^{(3)} = 0$. The age effects $\beta^{(k)}(x), k = 1,2,3$ are estimated (non-parameterically) from historical data, while the period and cohort effects are taken as stochastic processes. In the original proposal in [28], the period effect $\kappa^{(2)}$ is assumed to follow a random walk (i.e. unit root $AR(1)$ in discrete time),

$$\kappa^{(2)}(t) = \kappa^{(2)}(t-1) + \mu^{(2)} + \sigma^{(2)}\epsilon^{(2)},$$

where $\mu^{(2)}$ is the drift, $\sigma^{(2)}$ is the volatility, and $\epsilon^{(2)} \sim N(0,1)$ i.i.d. is the noise term. Alternatively, Cairns et al. [10] mention that $ARIMA$ models may provide a better fit, in particular fitting an $ARIMA(1,1,0)$ process for $\kappa^{(2)}$ based on 2007 CMI dataset.

For the cohort effect, Renshaw and Haberman [34] suggested using $ARIMA$ models for $\gamma^{(3)}(t-x)$; Cairns et al. [10] recommend the use of either $ARIMA(0,2,1)$ or $ARIMA(1,1,0)$. Renshaw and Haberman [34] and Cairns et al. [10] both assume $\gamma^{(3)}$ is independent of $\kappa^{(2)}$.

This model has identifiability issues, and one set of constraints could be

$$\sum_t \kappa^{(2)}(t) = 0, \quad \sum_x \beta^{(2)}(x) = 0, \quad \sum_{x,t} \gamma^{(3)}(t-x) = 0, \quad \text{and} \quad \sum_x \beta^{(3)}(x) = 1.$$

From a different perspective, Cairns, Blake, and Dowd [9] (CBD) proposed a model for $q(t,x) = 1 - P(Z(0);t,1,x)$, the probability of death in year t for someone aged x. Namely, they use

$$\operatorname{logit} q(t,x) = \beta^{(1)}(x)\kappa^{(1)}(t) + \beta^{(2)}(x)\kappa^{(2)}(t), \tag{M5}$$

where $\operatorname{logit}(y) = \log\left(\frac{y}{1-y}\right)$.

If we let n_a be the number of ages available in the data set for fitting, and take $x_{Ave} = n_a^{-1}\sum_i x_i$, the commonly used parameterization for the CBD model (M5) is

$$\beta^{(1)}(x) = 1, \qquad \text{and} \qquad \beta^{(2)}(x) = x - x_{Ave}. \tag{A.1}$$

Under these assumptions there are no identifiability issues.

Appendix B. Proofs of Analytic Estimates

Appendix B.1. Proof of Lemma 1.

Since the noise terms $\xi^{(k)}(u)$ are independent of $\kappa(s)$ for $u \neq s$, taking conditional expectation with respect to $Z(s) = \{\kappa^{(1)}(s), \xi^{(2)}(s)\}$, and writing in terms of the increments $\kappa(u) - \kappa(u-1)$

yields

$$\begin{aligned}\mathbb{E}\left[\kappa(t)-\kappa(s)\mid Z(s)\right] &= \sum_{u=s+1}^{t}\mathbb{E}\left[\kappa(u)-\kappa(u-1)\mid Z(s)\right] \\ &= \sum_{u=s+1}^{t}\mathbb{E}\left[\xi^{(1)}(u)+\xi^{(2)}(u)-\xi^{(2)}(u-1)\mid Z(s)\right]. \end{aligned} \tag{B.1}$$

By the independence assumption we have for $u \neq s+1$

$$\mathbb{E}\left[\xi^{(1)}(u)\mid Z(s)\right] = \mu^{(1)} \tag{B.2}$$

$$\mathbb{E}\left[\xi^{(2)}(u)-\xi^{(2)}(u-1)\mid Z(s)\right] = \mu^{(2)}p-\mu^{(2)}p = 0. \tag{B.3}$$

For $u = s+1$,

$$\mathbb{E}\left[\xi^{(2)}(s+1)-\xi^{(2)}(s)\mid Z(s)\right] = \mu^{(2)}p-\xi^{(2)}(s). \tag{B.4}$$

Combining (B.1)-(B.4), we obtain

$$\mathbb{E}\left[\kappa(t)\mid Z(s)\right] = \kappa(s)+(t-s)\mu^{(1)}+\mu^{(2)}p-\xi^{(2)}(s). \tag{B.5}$$

□

Appendix B.2. Proof of Lemma 2.

Since κ_1 has trend μ_1, $\mathbb{E}[\kappa_1(t)-\kappa_1(t-1)] = \mu_1$, and using conditional independence, we obtain,

$$\mathbb{E}\left[\kappa_1(T+t)\mid Z(T)\right] = \kappa_1(T)+\mu_1 t. \tag{B.6}$$

For the co-integration term $S(t)$, the expected values satisfy

$$\mathbb{E}[S(T+t)\mid Z(T)] = \mu_2+\phi\left(\mathbb{E}[S(T+t-1)\mid S(T)]-\mu_2\right). \tag{B.7}$$

The above gives a recursive equation for $t \mapsto \mathbb{E}[S(T+t)\mid Z(T)]$, with initial condition $\mathbb{E}[S(T+0)\mid Z(T)] = S(T)$, which can be solved to yield

$$\mathbb{E}[S(T+t)\mid Z(T)] = \mu_2(1-\phi^t)+\phi^t S(T). \tag{B.8}$$

Finally, using $\kappa_2(t) = \kappa_1(t)-S(t)$, and combining (B.6) with (B.8) leads to

$$\mathbb{E}[\kappa_2(T+t)\mid Z(T)] = \kappa_1(T)+\mu_1 t-\left(\mu_2(1-\phi^t)+\phi^t\left[\kappa_1(T)-\kappa_2(T)\right]\right).$$

as desired. □

Appendix B.3. Proof of Lemma 3

For $\xi^{(1)}(t,s)$, $\kappa^{(1)}$ is no different than a random walk with drift, so we have

$$\mathbb{E}[\kappa^{(1)}(t) \mid \kappa^{(1)}(s)] = \kappa^{(1)}(s) + \mu(t-s), \quad s \leq t.$$

Next, we take expectation on both sides of (42) to obtain the recursive relation

$$\mathbb{E}[\kappa^{(2)}(t) \mid Z(s)] = (1+\phi)\mathbb{E}[\kappa^{(2)}(t-1) \mid Z(s)] - \phi\mathbb{E}[\kappa^{(2)}(t-2) \mid Z(s)] \tag{B.9}$$

where $Z(s) = \{\kappa^{(1)}(s), \kappa^{(2)}(s), \kappa^{(2)}(s-1)\}$. Equation (B.9) is a recursive relation in t with general solution

$$\mathbb{E}[\kappa^{(2)}(t) \mid Z(s)] = c_1\phi^t + c_2, \tag{B.10}$$

where the constants c_1 and c_2 are to be determined. Plugging-in the initial conditions

$$\begin{aligned} c_1\phi^s + c_2 &= \mathbb{E}[\kappa^{(2)}(s) \mid Z(s)] = \kappa^{(2)}(s), \quad \text{and} && \text{(B.11)} \\ c_1\phi^{s+1} + c_2 &= \mathbb{E}[\kappa^{(2)}(s+1) \mid Z(s)] = (1+\phi)\mathbb{E}[\kappa^{(2)}(s) - \phi\kappa^{(2)}(s-1) \mid Z(s)] \\ &= (1+\phi)\kappa^{(2)}(s) - \phi\kappa^{(2)}(s-1). && \text{(B.12)} \end{aligned}$$

and solving for c_1, c_2 we obtain

$$c_1 = \phi^{1-s}\frac{\kappa^{(2)}(s) - \kappa^{(2)}(s-1)}{\phi - 1}, \qquad c_2 = \frac{\phi\kappa^{(2)}(s-1) - \kappa^{(2)}(s)}{\phi - 1}. \tag{B.13}$$

Finally, combining (B.13) with (B.10), we arrive at (44). $\square$